SEEDS: Some thoughts on Data Management for NASA missions.

Victor Zlotnicki
Earth and Space Sciences Division
and Physical Oceanography DAAC
Jet Propulsion Laboratory

Presented 7 February 2002, SEEDS Levels of Service Workshop, College Park, MD

1. A Program to ensure the generation and stewardship of data products, not just a Standards Maintenance Organization.

The old EOSDIS was a project conceived as responsible for downloading data from satellites, processing it into low and many higher levels of products, archiving all, and distributing them in small, manageable chunks to a very diverse clientele (scientists, students, lawyers, etc). All of this was centrally planned, funded, and implemented.

NewDISS, and now SEEDS emphasizes a heterogeneous, competitive environment of providers. We were called this week to comment of the levels of service that providers should provide, as drafted in "SEEDS, Requirements / LOS & Cost Model Working Paper, Current Draft, December 21, 2001" by G. Hunolt.

I believe another topic needs some debate before burying ourselves into a discussion over how many seconds or hours need to elapse between data arrival and availability: the need for a NASA ESE Data Processing and Management PROGRAM. In the SEEDS era, as was done before EOSDIS, data processing and management are funded by a variety of sources, each with their own constraints: the managers of satellite missions pay for some 'Mission Data Center' functions, (for ESSP missions, the proj. mgr. is replaced by the Principal Investigator), 'somebody' *should* pay for the 'Long term Archiving' function, perhaps a science NRA pays for the generation and distribution of some new data products (some 'Multimission' and perhaps some 'Science Data Center' functions). Perhaps an Applications NRA pays for the generation and distribution of products targeted to farmers, or fishermen,

In the late 1980s / early 1990s, and taking as example the Mission Data Center, the Project Mgr. paid for processing and for very limited distribution (to the science team). Wide distribution, reprocessing, etc were out of his scope. So was periodically (5 years or less) refreshing the tapes that held the data. This has not changed much in 2002: ESSPs with their tight budget for satellite, sensors, data system and data validation, and the need to plan for the inevitable hardware cost overruns, can be expected to pay for an extremely limited ground segment (both product generation and distribution). Similarly, why would the NASA HQ Manager of a particular science program, pay for data documentation, user support, SEEDS-wide technical exchanges, etc? Who will pay for infrastructure and its services. For the marginal cost of supporting users who had a 'standard format' imposed upon them? For periodically (again 5 years!) refreshing the media or data formats (in the not-so-unlikely event of standard data formats that do not survive the test of time).

A program needs to be in place, with a strong manager at NASA HQ or GSFC, to provide a forum to set priorities for data product generation, for data management including long term archiving, to review existing data management work, to set light-touch standards for the various data activities to satisfy, to guide new technology from other programs (ESTO for example) when it lowers cost and increases functionality for this program. Three key responsibilities of such a Program are (1) to articulate the needs and defend a budget for such activities, (2) to review the performance of the various data processing and management activities, killing if need be those who fail or whose job is essentially done, and (3) to maintain common standards. While maintaining standards is included, that is only part of this program's charter.

Such a program would, for example, identify that a particular ESSP makes under its funding a narrow range of products, but that NASA ESE goals to reach a wider community would benefit from products tailored to other communities. Such a program would fund and review long term archiving work, to ensure that the nation's investment in satellite systems is not falling off old magnetic tapes as happened so many times in the 1980s. And yes, such a program would maintain standards for data exchange and levels of service.

2. Levels of Service from a User's perspective.

Putting on one hat, that of a science *user*, I am mostly interested in the following:

Finding ***Suitable*** *Data.* Is there any estimate of parameter X at time Y in location Z? Is this accurate enough for me to bother spending time on it? This should, in the year 2001, be a non-problem: we have the GCMD, countless catalogs and inventories and projects to glue these, etc. But, these are only useful if data providers actually put information into them, and keep it updated, something that is lacking.

LOS:

1. parameter, time in months, location: planet Earth, data source (satellite, sensor who made it).
2. as 1, plus location to 10 degs, height by layer (upper ocean, lower troposphere), plus a few references to published papers that either made or used the data.
3. as 2, plus fine location and time (commensurate with data resolution), plus detailed algorithm and use explanations.

Notice that these are needed just to DECIDE whether the data are worth using. As a scientist tackling a new problem, any new dataset requires a time investment to explore, and identifying likely candidates, then eliminating 90% of those is the purpose of this step.

Accessing the data.

LOS, part 1:

1. I will receive a tape in a few weeks. I only need 10% of the tape's contents.
2. ftp download today. I only need 50% of the files' contents.
3. ftp download 'now'. The files contain exactly the data I need.

4. DODS or equivalent download now (does not make a 'disk file', but is read directly from my program). The data stream contains exactly the data I need.

Notice that for 2 and 3, 'today' and 'now' can be minutes to hours later. For example, a program can be spawned at the data center to select and subset the data, then put the files in some staging area. For 4 that is not acceptable, since my program will stop until the data are available.

LOS, part 2:

1. I can read the data within several days of receiving it
2. I can read the data within a few hours of receiving
3. I can read it as soon as I receive it.

Notice that a delay of several days can be caused either by poor format information, or by a cumbersome format and the lack of a reading program, such that writing readers is an onerous task, or by a standard format whose software does not properly install on my computer.

Using the data.

As data are used, peculiarities are found, and whether a scientist, a lawyer's technical aide or a legislative technical aide, it is necessary to have access to *very knowledgeable* people involved in the data production to find what the data are good for, and what they are not. Detailed documentation on algorithms and product generation is an alternative, longer path to the same answer.

LOS:

1. a few FAQs, a reference to a journal paper whose appendix describes the processing, or to papers that used the data
2. data center has (electronic) copies of technical reports detailing Algorithms, processing, etc
3. the data producer is alive and alert, user can contact her. Or, the data is old, but is *so well* documented that one can learn everything from the documentation.

3. A few Habits of Highly Effective Data Centers

I. The data are frequently exercised by very knowledgeable (science) users. The data center people are interested in their experience on accuracy/ usability/ applicability of the data. The data center reprocesses or causes others to reprocess data when problems are identified. The data center incorporates user comments into their FAQs on each data set. Sometimes this means having scientists on staff, or a postdoc program, or being co-located with many users (e.g., NCAR data section); sometimes it just means close collaboration between data center personnel and selected users.

II. The data center is closely linked to data producers. Whether the producers are the ground systems of large NASA projects, or 'Science Data Centers', or academic researchers who make data products as aids to scientific discovery. This closeness can be through co-location (e.g., Frank Wentz's www.ssmi.com; PO-DAAC at JPL near the Topex and Quikscat engineers; or Seawifs at GSFC), or simply close working interaction (e.g., NCAR data section with NCEP or ECMWF personnel).

III. Technology is used appropriately, to lower costs and increase functionality. For example, it is much cheaper to put 1 terabyte online (about $10K today, 2-3 times as much if fully redundant RAID) than to use tape robots; that requires buying just enough disk for the near term. It is much cheaper to deliver data electronically and automatically than to deliver media, especially if the media are tailor-made.

IV. The data center promptly incorporates user comments or complaints into their FAQs, and planned work on each data set (the difference between this and I is that this applies to general users, not just the most knowledgeable).

V. The 'level of service' is appropriate to the expected number of users. For example, there is no point in spending 200K to embellish a level 0 data set that can only be used by 2-3 people/groups worldwide. It is better for NASA to allocate those funds directly to one of the expected users.